HelioCAMPUS
Better Data. Brilliant Insights.

It's Not Magic

Understanding Data Science with Applications in Enrollment Management

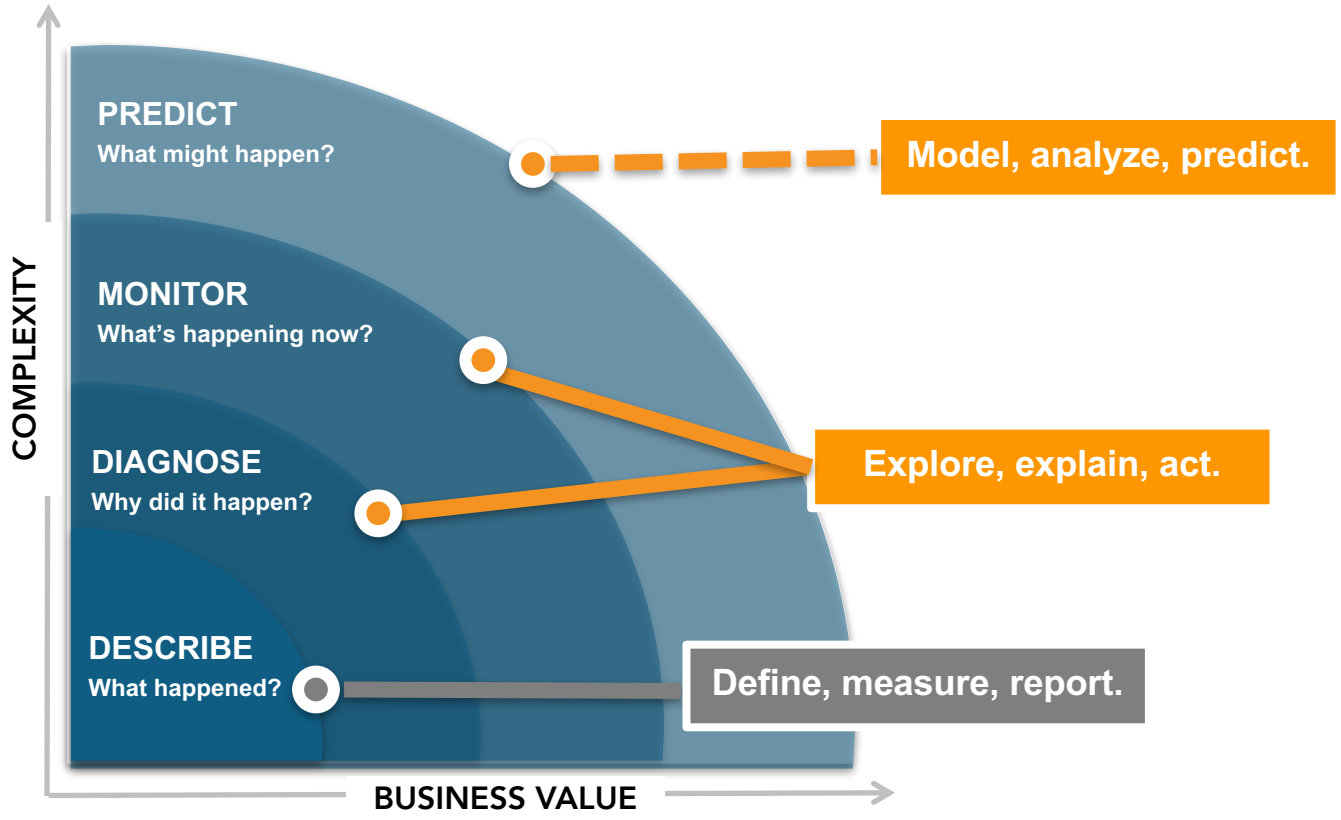
North Carolina Association for Institutional Research Conference 2019

Beyond the hype

Beyond the hype

- The hype…
    - Buzz about big data, artificial intelligence, machine learning, predictive analytics
- The reality…
    - Like any new technology, has its benefits and limitations
    - Can be a powerful tool when combined with organizational buy-in, knowledge and training

Data science or data analytics?



Why data science?

- Predict some future state or some current state that is unmeasurable
- Predictive can also be used to understand the “why” behind the what –
    - The model inputs are as important as the model outcome – are there hidden patterns that are visible when we control for other factors?

- Example: What are the common denominators behind students who have dropped out?

So you want to build a model

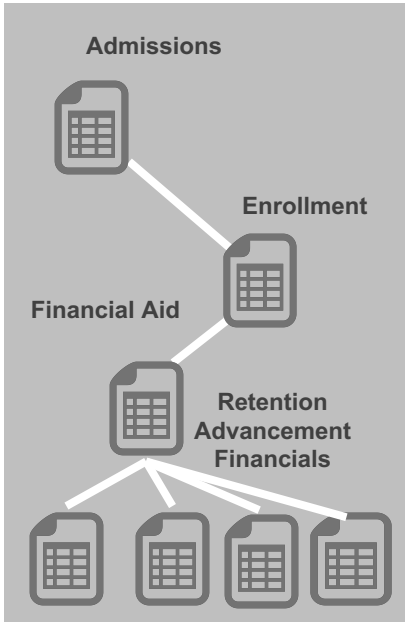
Data science project flow

How many new and returning students do we expect next term by academic program?

Which students are the most at risk for not returning next term?

How is financial aid and need related to yield at our institution?

Define Questions



Data Assembly

Exploration

Model Competition

- Random Forest
- K-Means Clustering
- Logistic Regression

Testing & Validation

Distribute Results

Predictive

HelioCAMPUS

Ask the right question

What is next year's enrollment going to be?

What is next year's enrollment going to be?

How many new students are enrolling next year?

How many students who are currently enrolled are going to come back?

What is next year's enrollment going to be?

How many new students are enrolling next year?

- Questions:
  - How many applications are we expecting?
  - If a given student applies, what is the likelihood that they will enroll?

How many students who are currently enrolled are going to come back?

- Questions:
  - Who is likely to graduate?
  - Who is likely to persist or drop out?

What is next year's enrollment going to be?

How many new students are enrolling next year?

- Questions:
  - How many applications are we expecting?
  - If a given student applies, what is the likelihood that they will enroll?

- Universe:
  - First time freshmen
  - Transfers
  - Certain majors/colleges

How many students who are currently enrolled are going to come back?

- Questions:
  - Who is likely to graduate?
  - Who is likely to persist or drop out?

- Universe:
  - Segmented by credit hours

Garbage in, garbage out

Data: the foundation of the model

How many new students are enrolling next year?

- Daily applications entered into the system
- Applicant-level data including HS academics, test scores, demographics

How many students who are currently enrolled are going to come back?

- Student-level data: credits, grades, demographics
- Historical datasets of previous students who were enrolled and did / did not re-enroll

Show me the magic

What is a model?

A **model** is a set of rules used to turn a set of inputs into an output.

An **algorithm** is how we come up with those rules.

What is a model?

Train the model:

$$algorithm(inputs) \rightarrow rules$$

Apply the model:

$$rules(inputs) \rightarrow output$$

Algorithms ahoy!

CLASSIFICATION

Enrollment Prediction
Identifying admitted students who are most likely to enroll

K-Nearest Neighbors
Random Forest

REGRESSION

Attribute Importance/ Influence on Retention
Understanding top predictors that correlate with retention

Logistic Regression
Linear Regression

CLUSTERING

Student Segmentation
Finding related sub-populations of students

K-Means
Hierarchical Clustering

DIMENSIONALITY REDUCTION

Simplifying and Combining Attributes
Discovering correlated attributes and streamlining analyses

Randomized PCA
Kernel Approximation

Modeling re-enrollment likelihood

Inputs:

- Independent variables: student's cumulative GPA, cumulative credits, total dropped classes, full or part time, financial aid status, number of previous terms enrolled
- Dependent variable: whether the student re-enrolled

Algorithm:

- Elastic net regression

Output:

- 0 to 1 "score"

Measure twice, cut once

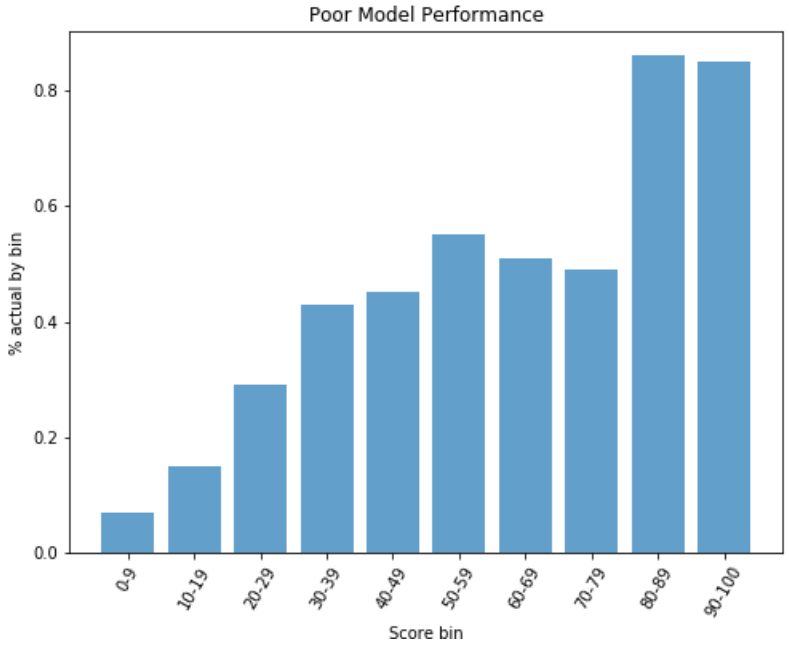
How do we know it works?

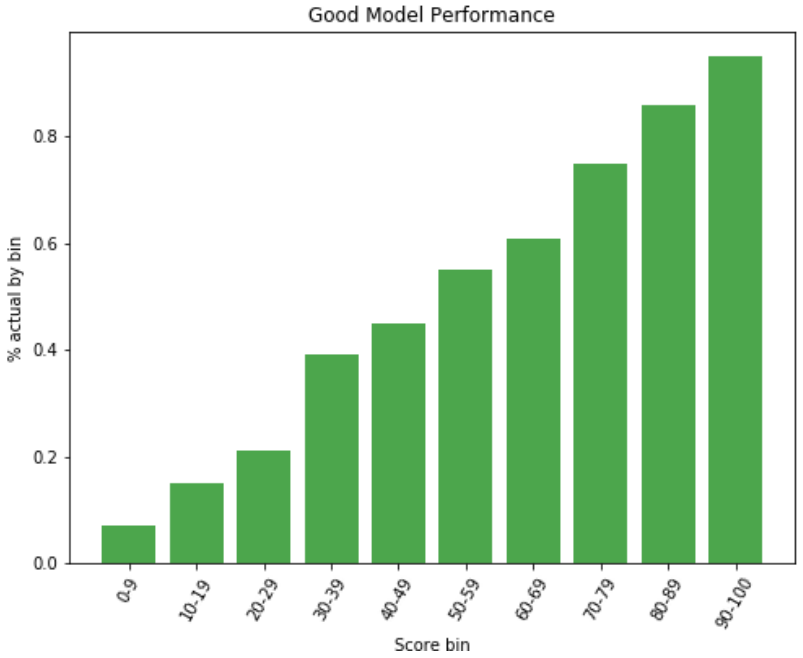
- Evaluate the model:

$$algorithm(test\ inputs) \rightarrow output$$

$$model\ output \sim actual\ output$$

How do we know it works?

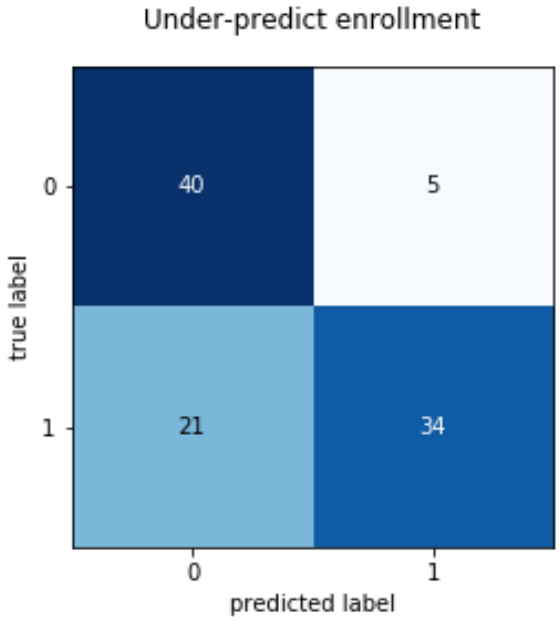


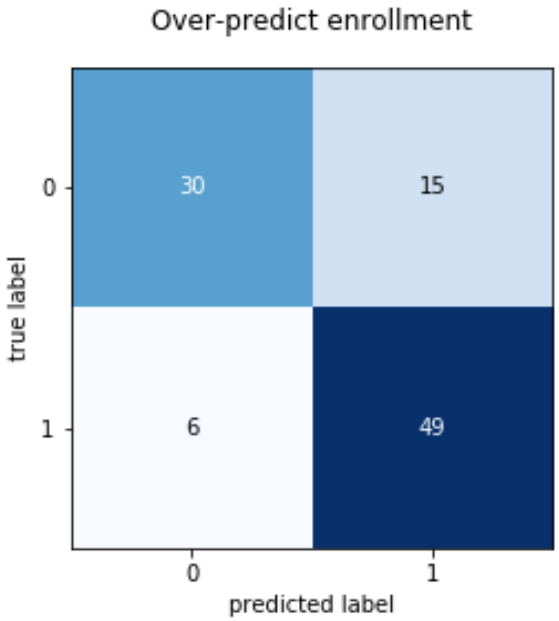


22

HelioCAMPUS

How do we know it works?





Showtime

How are we going to use it?

- Build out infrastructure
    - Table inside a SQL database
    - Script that runs regularly to refresh the model
- Train and deploy to end users
    - Dashboard or other front-end tool
    - Documentation and training materials

Questions